

# GENERALIZATIONS RELATED TO HYPOTHESIS TESTING WITH THE POSTERIOR DISTRIBUTION OF THE LIKELIHOOD RATIO.

I. SMITH

*Laboratoire des Sciences du Climat et de l'Environnement ; IPSL-CNRS, France.  
Université de Nice Sophia-Antipolis, CNRS, Observatoire de la Côte d'Azur, France.*

A. FERRARI

*Université de Nice Sophia-Antipolis, CNRS, Observatoire de la Côte d'Azur, France*

**ABSTRACT.** The Posterior distribution of the Likelihood Ratio (PLR) is proposed by Dempster in 1974 for significance testing in the simple vs composite hypotheses case. In this hypotheses test case, classical frequentist and Bayesian hypotheses tests are irreconcilable, as emphasized by Lindley's paradox, Berger & Selke in 1987 and many others. However, Dempster shows that the PLR (with inner threshold 1) is equal to the frequentist p-value in the simple Gaussian case. In 1997, Aitkin extends this result by adding a nuisance parameter and showing its asymptotic validity under more general distributions. Here we extend the reconciliation between the PLR and a frequentist p-value for a finite sample, through a framework analogous to the Stein's theorem frame in which a credible (Bayesian) domain is equal to a confidence (frequentist) domain.

This general reconciliation result only concerns simple vs composite hypotheses testing. The measures proposed by Aitkin in 2010 and Evans in 1997 have interesting properties and extend Dempster's PLR but only by adding a nuisance parameter. Here we propose two extensions of the PLR concept to the general composite vs composite hypotheses test. The first extension can be defined for improper priors as soon as the posterior is proper. The second extension appears from a new Bayesian-type Neyman-Pearson lemma and emphasizes, from a Bayesian perspective, the role of the LR as a discrepancy variable for hypothesis testing.

## 1. INTRODUCTION

**1.1. Classical hypotheses test methodologies.** Simple versus composite hypotheses testing is a general statistical issue in parametric modeling. It consists for a given observed dataset  $x$  in choosing among the hypotheses

$$(1) \quad H_0 : \theta = \theta_0 \quad H_1 : \theta \in \Theta_1$$

where the distribution of  $x$  is characterized by the underlying unknown parameter  $\theta$ . Under the alternative hypothesis  $H_1$ ,  $\theta$  takes a value different from the point  $\theta_0$ , and the uncertainty of  $\theta$  is described by a prior probability density function  $\pi_1(\theta)$  which is positive only for  $\theta \in \Theta_1$ . We assume that the data model  $p(x|\theta)$  has the same expression under  $H_0$  and  $H_1$ .

To choose among  $H_0$  and  $H_1$ , a test statistic  $T(x)$  (such as the Generalized Likelihood Ratio) is generally compared to a threshold  $\zeta$  and one decides to choose  $H_0$  if  $T(x)$  is greater than  $\zeta$ . If  $H_1$  is chosen whereas the true underlying  $\theta$  was equal to  $\theta_0$ , a type I error is made in the decision. Under the classical Neyman paradigm (see Neyman and Pearson (1933); Neyman (1977)), the

---

*E-mail addresses:* zazoo@mac.com, andre.ferrari@unice.fr.

*Key words and phrases.* hypothesis testing, PLR, p-value, likelihood ratio, frequentist and Bayesian reconciliation, Lindley's paradox, invariance, Neyman-Pearson lemma.

Part of this work was published in Smith and Ferrari (2014).

threshold  $\zeta$  is chosen so that the probability of the type I error lies under (or is equal to) some fixed level  $\alpha$ , typically a 5% error rate. Instead of inverting this function, a p-value can be defined in order to serve as the test statistic to be directly compared to the 5% level (Lehmann and Romano (2005)):

$$(2) \quad p_{\text{val}}(T(x_0)) = \Pr(T(x_0) < T(x)|\theta_0)$$

where  $x_0$  is the observed dataset and  $x$  the variable of integration. Note that with this notation,  $H_0$  is rejected when  $p_{\text{val}}(T(x_0))$  is greater than some threshold.

On the Bayesian side, the test statistic classically used (Robert (2007)) is the Bayes Factor (BF) defined by

$$\text{BF}(x) = \frac{p(x|\theta_0)}{\int d\theta p(x|\theta)\pi_1(\theta)}$$

Making a binary decision consists of choosing  $H_0$  if  $\text{BF}(x)$  is greater than some threshold, and the choice of the threshold is made in general by a straight interpretation of the BF. The Jeffreys's scale for example states that if the observed BF is between 10 and 100 there is a strong evidence in favor of  $H_0$ . The mere posterior probability  $\Pr(H_i|x)$  of an hypothesis may also be considered by itself.

A practical issue of the BF in the simple vs composite hypotheses test is that it is defined up to a multiplicative constant if the prior  $\pi_1$  is improper<sup>1</sup> even though the posterior distribution is proper. Partial BFs account for this issue by somehow using part of the data to update the prior into a proper posterior, and then use this posterior as the prior for the rest of the data. The most simply defined Partial BF is the Fractional BF (FBF) proposed by O'Hagan (1995).

A related and more fundamental issue is Lindley's paradox, initially studied by Jeffreys (1961) and called a paradox by Lindley (1957), which shows among others that, when testing a simple vs a composite hypothesis, the null hypothesis  $H_0$  is too highly favoured against  $H_1$  for a natural diffuse prior under  $\Theta_1$ . More precisely, for example in the test of the mean of a Gaussian likelihood, the p-value  $|x|$  defines the uniformly most powerful test, which is a very strong optimal property even according to at least part of the Bayesian community. However, for a fixed prior and some dataset  $x$  that adjusts so that the associated classical p-value remains fixed (so that the evidence for  $H_0$  shall not change),  $\Pr(H_0|x)/\Pr(H_1|x)$  tends to 1 as the sample size increases. This issue, intensively discussed and developed (see Tsao (2006) for a quite recent study), is consensually considered as a real trouble by a quite large part of the community. Unlike the BF, other tests like the FBF or the Bernardo (2011) test do not suffer from this problem in Lindley's frame. Other ideas have been developed which prevent Lindley's frame from occurring, avoiding troubles for the BF. Berger and Delampady (1987) for example argue that testing a simple hypothesis is an unreasonable question. Some other references will be given in the section 2.1.

Among many frequentist and Bayesian p-values (several are listed by Robins et al. (2000)), the next most classical Bayesian-type hypotheses test statistic is the posterior predictive p-value, highlighted by Meng (1994). Unlike the BF which only integrates over the parameter space  $\Theta$ , the posterior predictive p-value integrates over the data space  $\mathcal{X}$ , like frequentist p-values. But unlike the frequentist p-value which integrates under the frequentist likelihood  $p(x|\theta_0)$ , it integrates under the predictive likelihood  $p(x^{\text{pred}}|x_0) = \int d\theta p(x^{\text{pred}}|\theta)\pi(\theta|x_0)$  where  $x_0$  is the observed dataset. In a frequentist p-value only a statistic (ie a function of  $x$  only) can define the domain of integration. On the contrary, in the posterior predictive p-value, a discrepancy variable (function of both  $x$  and  $\theta$ ) can be used to define the domain of integration. Note that the choice of the discrepancy variable to use there remains an issue.

Although a bit less classical, the approach of Evans (1997) needs to be introduced because the tool and some of its properties are interesting and closely related to the ones derived in this paper. In the simple vs composite test case presented up to now, the tool proposed by Evans

---

<sup>1</sup> $\pi_1$  is called improper if its integral over  $\Theta_1$  is infinite, which occurs if  $\pi_1$  is constant over an unbounded domain for example.

(1997) and the ones studied in this paper are even mathematically equal. But the tool proposed by Evans (1997) is defined to test more generally  $H_0 : \Psi(\theta) = \psi_0$  for a parameter of interest  $\psi = \Psi(\theta)$ . The test statistic consists of measuring the Observed Relative Surprise (ORS) related to the hypotheses by computing:

$$(3) \quad \text{ORS}(x) = \Pr \left( \frac{\pi_{\Psi}(\Psi(\theta)|x)}{\pi_{\Psi}(\Psi(\theta))} \geq \frac{\pi_{\Psi}(\psi_0|x)}{\pi_{\Psi}(\psi_0)} \middle| x \right)$$

The relative belief ratio of  $\psi$  defined by  $\text{RB}(\psi) = \pi_{\Psi}(\Psi(\theta)|x)\pi_{\Psi}(\Psi(\theta))^{-1}$  is measuring the change in belief in  $\psi$  being the true value of  $\Psi(\theta)$  from *a priori* to *a posteriori*. So if  $\text{RB}(\psi_0) > 1$  we have evidence in favor of  $H_0$ . Relative belief ratios are discussed in Baskurt and Evans (2013) where  $\text{RB}(\psi_0)$  is presented as the evidence for or against  $H_0$  and (3) is presented as a measure of the reliability of this evidence. This leads to a possible resolution of Lindley's paradox as the relative belief ratio can be large and ORS small without contradiction. See Example 4 of Baskurt and Evans (2013) and note that Evans (1997) shows that ORS converges to the classical p-value as the prior becomes more diffuse in this example.

**1.2. Posterior distribution of the Likelihood Ratio (PLR).** Let's focus again on the simple vs composite hypothesis test. Contrary to the posterior predictive p-value, the Posterior distribution of the Likelihood Ratio (PLR) does not integrate over some data which are unobserved, but only integrates over  $\Theta$ . It still conditions upon the only observed variable, namely  $x_0$ , like for the BF, but on a domain defined from a divergence variable, like the posterior predictive p-value. This statistic proposed by Dempster (1973) is defined by

$$(4) \quad \text{PLR}(x, \zeta) = \Pr(\text{LR}(x, \theta) \leq \zeta | x)$$

where  $\text{LR}(x, \cdot)$  is the Likelihood Ratio

$$\text{LR}(x, \theta) = \frac{p(x|\theta_0)}{p(x|\theta)} \quad \theta \in \Theta_1$$

Since  $\theta$  is random, the deterministic function  $\text{LR}(x, \cdot)$  evaluated at the random variable  $\theta$  becomes naturally random with some posterior distribution characterized by its cumulative distribution, the PLR. As emphasized by Birnbaum (1962), Dempster (1973) and Royall (1997), the threshold  $\zeta$  which compares the original likelihoods under  $H_0$  and under  $H_1$  is directly interpretable and can be chosen the same way an error level  $\alpha$  is chosen in the Neyman-Pearson paradigm. “ $\text{PLR}(x, 1) = 0.1$ ” for example reads “The probability that the likelihood of  $\theta_1$  is more than the likelihood of  $\theta_0$  is 0.1.”

The PLR can therefore be used for a binary decision, by fixing  $\zeta$  and deciding to reject  $H_0$  if  $\text{PLR}(x, \zeta)$  is greater than, say, 0.9. One can check if the binary decision is sensitive to the choice of both thresholds by making the test for several thresholds and see if the decision is different. In the extreme case, note that due to the nice definition of the PLR, one can simply display  $\text{PLR}(x, \zeta)$  as a function of  $\zeta$  to get a broad view. The range of  $\zeta$  under which  $\text{PLR}(x, \zeta)$  grows typically from 0.2 to 0.8 indicates if the decision for  $H_0$  or  $H_1$  is clear, or not. As soon as the posterior can be sampled, these computations and graphs are very easy to display as will be explained later.

The PLR has been first proposed by Dempster (1973, 1997), then studied especially by Aitkin (1997) and Aitkin (2010) but also used and analyzed by Aitkin et al. (2005, 2009). As mentioned in the previous subsection, it turns out that the PLR is also closely related to the ORS proposed by Evans (1997), which generalizes the PLR. The PLR is also closely related to the e-value associated to the Full Bayesian Significance Test (FBST) from Pereira and Stern (1999) and slightly revisited by Borges and Stern (2007) which then somehow generalizes the PLR by adding a reference distribution on  $\theta$ , and by systematically dealing with the case where the null hypothesis domain  $\Theta_0$  has a dimension less than  $\Theta_1$  but which is not necessarily restricted to the point  $\Theta_0 = \{\theta_0\}$ . We do not list the results found by these different analyses, apart from some specifically mentioned ones.

The PLR turns out to be a natural Bayesian measure of evidence of the studied hypotheses since it involves only the posterior distribution of  $\theta$  (no integral over  $\mathcal{X}$ ) and the likelihood, claimed by Birnbaum (1962), Royall (1997) and others, to be the only tool that can measure evidence. Unlike the BF, the PLR is well defined for an improper prior as soon as the posterior is proper, and is not subject to Lindley’s paradox. It is also invariant under any isomorphic transformation of the  $\mathcal{X}$  space and any transformation of the  $\Theta$  space, as a consequence of being a mere function of the likelihood. These last properties were emphasized for example for the e-value associated to the FBST.

The PLR also consists in a natural alternative to the BF in different regards. To start with, the PLR first compares (compares  $p(x|\theta_0)$  and  $p(x|\theta)$ ) and then integrates, whereas the BF first integrates and then compares (compares  $p(x|\theta_0)$  and  $\int d\theta p(x|\theta)\pi(\theta)$ ). Second, Newton and Raftery (1994) and many others show that if the prior under  $H_1$  is proper, the BF is simply the posterior mean of the LR, ie the mean of the distribution described by the PLR<sup>2</sup>. However a point estimate is in general not given alone but accompanied by an uncertainty indicator. Smith and Ferrari (2010) show that the posterior mean of the LR raised at some power is equal to the FBF introduced previously; the mean of the PLR is given by the BF and its variance is easily related to the FBF. However, Smith (2010) shows that the Generalized Likelihood Ratio bounds the support (values of  $\zeta(x)$  for which  $\text{PLR}(x, \zeta) > 0$ ) of the PLR and that at this lower bound the PLR in general starts by an infinite derivative. In addition to this theoretical result, numerical examples also indicate that the posterior density function of the LR is in general highly asymmetric. Therefore, the BF (point estimate of the LR) or any standard centered credible intervals do not appear to be relevant inferences about the LR seen as random variable. Instead, the same way the BF is to be thresholded, the actual information about  $\text{LR}(x, \theta)$  which seems to be relevant, and invariant under the transformation  $\text{LR}(x, \theta) \mapsto (\text{LR}(x, \theta))^{-1}$ , is to indicate its cumulative posterior distribution, which is precisely the PLR.

In practice, the PLR can be straightforwardly computed as soon as the posterior distribution  $\pi(\theta|x)$  is sampled. Just obtain from a Monte Carlo Markov Chain (MCMC) algorithm an almost i.i.d. chain  $\{\theta^{[1]}, \dots, \theta^{[m]}\}$  from the posterior distribution  $\pi(\theta|x)$  and compute  $\text{LR}(x, \theta^{[i]})$  for each sample. The resulting histogram sketches the posterior density of the LR and the plot of the empirical cumulative distribution of the LR chain sketches the PLR as a function of  $\zeta$ .

The PLR has been realistically and thoroughly applied by Smith (2010) to the detection of extra-solar planets from images acquired with the dedicated instrument SPHERE mounted on the Very Large Telescope. At this moment, only very finely simulated images were available. The PLR has been applied to two simulated datasets, one in which no extra-solar planet is present (dataset simulated under  $H_0$ ) and the other in which an extra-solar planet is present ( $H_1$  dataset). Although the extra-solar planet is very dark ( $10^6$  times less bright than the star it surrounds), close to the star (angular distance in the sky of 0.2 arcseconds i.e.  $6.10^{-5}$  degrees), and although only  $2 \times 20$  images were used, thanks to the quality of the optical instruments and of the statistical model the detection and not detection were evident, with  $\text{PLR}(x, 0.1) = 0.0$  for the dataset under  $H_0$  and  $\text{PLR}(x, 0.1) = 0.94$  for the dataset under  $H_1$ . As studied by Smith (2010), the statistical model and consecutive method are very satisfying compared to classical methods.

**1.3. Problematics addressed here.** Despite its potential interest the PLR has not been extensively studied up to now. This paper aims at contributing in this investigating work by some new results.

In the simple vs composite hypotheses test case, it turns out that the PLR plays a strong role in understanding the possible reconciliation between frequentist and Bayesian hypothesis testing. The PLR with inner threshold  $\zeta = 1$  is simply equal to some frequentist p-value for some “likelihood - prior - hypotheses” combinations. Dempster (1973); Aitkin (1997) have first

---

<sup>2</sup>Alternatively, note that if we had defined the BF and LR with the alternative hypothesis at the numerator of these fractions, the BF would have been the prior mean of the LR.

noticed and highlighted this equivalence when testing the mean of a gaussian likelihood with a uniform prior.

In the section 2, we extend the conditions of this equivalence result under a frame analogous to the one used to reconcile confidence and credible domains. The subsection 2.1 synthesizes the long quest of reconciliation between frequentist and Bayesian hypotheses tests, the subsection 2.2 proves and discusses the reconciliation reached between the PLR and some frequentist p-value in such an invariant frame, the subsection 2.3 gives examples and perspectives, and the subsection 2.4 discusses the connection between this reconciliation result and the one obtained between (frequentist) confidence domains and (Bayesian) credible regions.

Aitkin (1997) and Aitkin (2010) extended the PLR definition to an hypotheses test frame identical to the one presented at the end of the subsection 1.1, namely  $H_0 : \Psi(\theta) = \psi_0$ , also considered by Evans (1997) and others. However, the PLR has not been yet generalized to the general composite vs composite hypotheses test. The generalization is somehow unnatural for a frequentist p-value because for a simple hypothesis  $H_0 : \theta = \theta_0$  the p-value is a frequentist probability conditioned on the fixed parameter  $\theta_0$  (see equation 2), although a conditional probability cannot be defined on a composite set  $\Theta_0$  if no probability distribution over  $\Theta$  is used. By contrast, the PLR is reciprocally a probability conditioned upon the observed dataset  $x_0$ , and  $x_0$  naturally remains fixed under a composite hypothesis. Therefore, the transition from simple to composite null hypothesis does not raise immediate obstacles for the PLR. However, a joint measure on the parameter spaces of both hypotheses is still required.

The section 3 proposes and motivates two generalizations of the PLR. The mathematical expressions of the two extensions are simply given and rephrased in the subsection 3.1. The first extension in particular enables the use of improper priors as soon as the posterior is proper. It can therefore be used in the subsection 3.2 for the detection of precipitation change where an almost improper prior is to be used but leads to a proper posterior. On the other side, the second extension, made of two symmetrical probabilities, appears in a Bayesian version of the Neyman-Pearson lemma. As detailed in the subsection 3.3, the two joint measures associated to no specific discrepancy variable lead through the lemma to the discrepancy variable  $LR(x_0, \theta)$ .

A concluding discussion is proposed in the section 4. The appendices essentially present the proofs of the mathematical results.

## 2. EQUIVALENCE BETWEEN THE PLR AND A FREQUENTIST P-VALUE

**2.1. Previous tentative reconciliations of frequentist and Bayesian tests.** As introduced in the section 1.1, Lindley’s paradox presents a frame where  $\Pr(H_0|x)$  (often thought as being *the* Bayesian measure of evidence) may be expected to be equal to the frequentist p-value, but happens not to be. Also, the BF is not satisfying in the frame “point null hypothesis  $H_0$  and diffuse prior  $\pi_1$ ”. This highlights the need for other Bayesian-type hypotheses tests, but also raises more generally the question of reconciliation between frequentist and Bayesian hypotheses tests.

The conditions upon which frequentist (Neyman (1977)) and Bayesian (Jeffreys (1961)) answers agree is always of interest in order to understand the interpretation of the procedures and the limits of the two paradigms, somehow defined by what they are *not*.

A first approach to see when could frequentist and Bayesian hypotheses tests be unified consists of analyzing, for different hypotheses likelihoods and priors, when are the classical p-value and  $\Pr(H_0|x)$  equal. These two concepts are to be compared because they both seem to handle only  $H_0$  and in very simple ways, one from the frequentist the other from the Bayesian perspectives<sup>3</sup>. It turns out that unlike for a composite null hypothesis (e.g. Casella and Berger (1987)), for a point null hypothesis Lindley’s paradox  $\Pr(H_0|x) > p_{\text{val}}(x)$  always seems to hold. Berger and Sellke (1987) in particular show that among very broad classes of priors  $\Pr(H_0|x) > p_{\text{val}}(x)$  always holds for  $\Pr(H_0) = 0.5$ . Also see the extensive list of references included. Oh and DasGupta (1999) follows this analysis by studying the effect of the choice of  $\Pr(H_0)$ .

<sup>3</sup>Note however that  $H_1$  is implicitly taken into account through the marginal distribution of  $x$  in  $\Pr(H_0|x)$ .

Another approach consists of modifying the standard frequentist procedure and/or the standard Bayesian hypotheses test procedure, but still relying on the p-value and  $\Pr(H_0|x)$ , to see if they can then be made equivalent. Berger and Delampady (1987) for example study “precise” (concentrated) but not exactly “point” hypotheses, Berger et al. (1994) use frequentist p-values computed from a likelihood conditioned upon a set in which lies the observed dataset, not on the dataset itself, and define a non-decision domain in the BF test procedure. Sellke et al. (2001) advocate calibrating (*rescaling*) the frequentist p-value to relate this new statistic to other test statistics.

As already mentioned in the section 1.1, one can also try to unify the p-value to Bayesian type statistics fully *different* from the BF, to see when frequentist and Bayesian types hypotheses tests can be made equivalent. In particular, when Dempster (1973) proposed to use the PLR, he also mentioned that when testing the mean of a normal distribution, the PLR is equal to the classical frequentist p-value when computed for a uniform prior and with inner parameter  $\zeta = 1$ . This fundamental result was again emphasized by Aitkin (1997) and Dempster (1997).

Aitkin (1997) asymptotically extended this result to any regular distribution, making use of the asymptotic convergence of a regular distribution towards a normal distribution. For any regular continuous distribution and a smooth prior, the PLR, with  $\zeta = 1$ , tends asymptotically to the classical p-value. Also, with a nuisance parameter  $\eta$  and still calling  $\theta$  the tested parameter, he defines LR by  $\text{LR}(x, \theta, \eta) = p(x|\theta_0, \eta)/p(x|\theta, \eta)$ , in which case under the same conditions as in the previous case the PLR is equal to a p-value. For a normal distribution, when testing the mean and considering the variance as a nuisance parameter, the result is also true for a finite sample.

**2.2. New reconciliation result.** The sets of conditions found by Dempster (1973) and Aitkin (1997) under which the PLR (with  $\zeta = 1$ ) is equal to a p-value are directly related to the test of the mean of a normal distribution under a uniform prior. The next subsection generalizes this exact finite-sample result under the frame of statistical invariance. As will be discussed at the end of the section, although the technical conditions derived here may be relaxed, it may be difficult to find, at least within the current statistical frame, a fundamentally more general frame of conditions for an equality between the PLR and a p-value to hold.

As presented in current classical textbooks in Bayesian statistics (Berger (1985), Robert (2007)), invariance in statistics arises from the invariant Haar measure defined on some topological group. Throughout this subsection and the related appendices, we will use the notions and results synthesized by Nachbin (1965) and Eaton (1989). The tools necessary to understand the result are introduced in the appendix 1.

In this frame, the PLR (given by an integral over the parameter space  $\Theta$ ) can be reexpressed as an integral over the sample space  $\mathcal{X}$ , equal to a p-value for  $\zeta = 1$ . In this subsection  $x$  and  $\theta$  denote random variables or variables of integration according to the context.

First, for clarity, we give the equivalence between the PLR and a frequentist integral under the assumption that the sample space  $\mathcal{X}$ , the parameter space  $\Theta$  and the transformations group  $\mathcal{G}$  are isomorphic.

**Theorem 1.** *Call  $\mathcal{P}_\Theta = \{p(\cdot|\theta), \theta \in \Theta\}$  a family of probability densities with respect to the Lebesgue measure on  $\mathcal{X}$ , and call  $\mathcal{G}$  a group acting on  $\mathcal{X}$ . Assume that  $\mathcal{P}_\Theta$  is invariant under the action of the group  $\mathcal{G}$  on  $\mathcal{X}$  and note  $\bar{g}\theta$  the induced action of the element  $g \in \mathcal{G}$  on the element  $\theta \in \Theta$ . Call  $H^r$  and  $H^l$  respectively a right and left Haar measures of  $\mathcal{G}$  and assume that*

- (1)  $\mathcal{G}$ ,  $\mathcal{X}$  and  $\Theta$  are isomorphic.
- (2) The prior measure  $\Pi^r$  is the measure induced by  $H^r$  on  $\Theta$ .
- (3) The measure induced by  $H^l$  on  $\mathcal{X}$  is absolutely continuous with respect to the Lebesgue measure. Call  $\pi^l$  the corresponding density.
- (4) The marginal density of  $x$  is finite, so that the posterior measure  $\Pi_x^r$  on  $\Theta$ , classically defined by the equation (23), defines the posterior probability  $\Pr(\cdot|x_0)$ .

Then, the PLR defined by the equation (4) can be reexpressed for any  $\zeta > 0$  as the frequentist integral:

$$(5) \quad \text{PLR}(x_0, \zeta) = \Pr \left( \frac{p(x_0|\theta_0)}{\pi^l(x_0)} \leq \zeta \frac{p(x|\theta_0)}{\pi^l(x)} \mid \theta_0 \right)$$

where  $x_0 \in \mathcal{X}$  is the observed data and  $\theta_0 \in \Theta$  the parameter value under the null hypothesis.

A more general theorem (Theorem 2) derived in a frame which avoids the Lebesgue assumption and may involve more technical conditions is proved in Appendix 2. Theorem 1 is a consequence of Theorem 2 and its proof is given in Appendix 3.

The assumption that  $\mathcal{G}$  and  $\mathcal{X}$  are isomorphic is easily relaxed by replacing the sample space by the space of a sufficient statistic. Recall that if  $X$  is a random variable whose probability distribution is parametrized by  $\theta$ ,  $S(X)$  is called a sufficient statistic of  $\theta$  if the probability distribution of  $X$  conditioned upon the random variable  $S(X)$  does not depend on  $\theta$ . Note that according to the Darmois (1935) theorem, among families of probability distributions whose domains do not vary with the parameter being estimated, only in exponential families is there a sufficient statistic whose dimension remains bounded as the sample size increases.

The expression of the theorem 2 is simply extended by replacing  $X$  by a sufficient statistic  $S(X)$  in the assumptions and by replacing in the frequentist integral the probability density of  $X$  by the one of  $S(X)$ :

**Corollary 1.** Call  $\mathcal{P}_\Theta = \{p(\cdot|\theta), \theta \in \Theta\}$  a family of probability densities with respect to any measure on  $\mathcal{X}$ . Call  $S(X)$ , for  $X \in \mathcal{X}$ , a sufficient statistic of  $\theta$  and  $\mathcal{P}_{S,\Theta} = \{p_S(\cdot|\theta), \theta \in \Theta\}$  the family of probability densities of  $S(X)$  with respect to the Lebesgue measure on  $S(\mathcal{X})$ . Call  $\mathcal{G}$  a group acting on  $S(\mathcal{X})$ . Assume that  $\mathcal{P}_{S,\Theta}$  is invariant under the action of the group  $\mathcal{G}$  on  $S(\mathcal{X})$  and note  $\bar{g}\theta$  the induced action of the element  $g \in \mathcal{G}$  on the element  $\theta \in \Theta$ . Call  $H^r$  and  $H^l$  respectively any right and left Haar measures of  $\mathcal{G}$ . Assume that

- (1)  $\mathcal{G}$ ,  $S(\mathcal{X})$  and  $\Theta$  are isomorphic.
- (2) The prior measure  $\Pi^r$  is the measure induced by  $H^r$  on  $\Theta$ .
- (3) The measure induced by  $H^l$  on  $S(\mathcal{X})$  is absolutely continuous with respect to the Lebesgue measure. Call  $\pi^l$  the corresponding density.
- (4) The marginal density of  $x$  is finite, so that the posterior measure  $\Pi_x^r$  on  $\Theta$ , classically defined by the equation (23), defines the posterior probability  $\Pr(\cdot|x_0)$ .

Then, the PLR defined by the equation (4) can be reexpressed, with  $x_0 \in \mathcal{X}$ ,  $\theta_0 \in \Theta$  and  $\zeta > 0$ , as the frequentist integral:

$$(6) \quad \text{PLR}(x_0, \zeta) = \Pr \left( \frac{p_S(S(x_0)|\theta_0)}{\pi^l(S(x_0))} \leq \zeta \frac{p_S(S(x)|\theta_0)}{\pi^l(S(x))} \mid \theta_0 \right)$$

where  $x_0 \in \mathcal{X}$  is the observed data and  $\theta_0 \in \Theta$  the parameter value under the null hypothesis.

The proof follows the proof of the theorem 1 in the Appendix 3.

By evaluating  $\zeta = 1$  in the result, the PLR with  $\zeta = 1$  is easily and finally shown to be equal to a frequentist p-value, where the test statistic is a weighted marginal likelihood of the sufficient statistic  $S(x)$ .

**Corollary 2.** Under the assumptions of the corollary 1, the PLR with inner threshold  $\zeta = 1$  is equal to a p-value:

$$(7) \quad \text{PLR}(x_0, 1) = p_{\text{val}}(T(x_0))$$

with the test statistic

$$(8) \quad T(x) = \frac{p_S(S(x)|\theta_0)}{\pi^l(S(x))}$$

The corollary 2 can be reexpressed as the fact that under the invariance assumptions, rejecting  $H_0$  when  $\text{PLR}(x_0, 1) > p$  is equivalent to rejecting  $H_0$  when  $p_{\text{val}}(T(x_0)) > p$  where the p-value

is based on the idea of rejecting  $H_0$  when  $T(x_0)$  defined in equation (8) (observed weighed likelihood under  $H_0$ ) is not large enough.

**2.3. Examples and perspective.** Dempster (1973) has shown that the PLR is equal to the classical p-value associated to the test statistic  $T(x) = |\bar{x} - \theta_0|$  when testing the mean of a normal family for  $X$  with a uniform prior on  $\Theta$ . The corollary 2 extends this result since the normal family is one of the distributions invariant under translation when testing the location parameter, the uniform prior (i.e. Lebesgue measure) is the measure induced from the right Haar measure associated to translation, and the test statistic  $T(\cdot)$  is a monotone function of  $p_S(S(\cdot)|\theta_0)\pi^l(S(\cdot))^{-1}$  since the translation (sum) is commutative, so that  $\Delta(g) = 1$  for all  $g \in \mathcal{G}$  and so  $\pi^l$  is constant.

The result proved here concerns all distributions invariant under some group transformation, under the assumptions that there exists a sufficient statistic and that the sets  $\mathcal{G}$ ,  $S(\mathcal{X})$  and  $\Theta$  are isomorphic. Assume for example that the likelihood  $p_S$  has the typical form  $p_S(S(x)|\theta) = \theta^{-1}f(S(x)\theta^{-1})$ . The likelihood is invariant under the scale transformation  $g(S(x)) = \alpha \times S(x)$  and the actions on  $S(\mathcal{X})$  and  $\Theta$  are identical. Note that  $Uf(U)$  with  $U = S(X)\theta^{-1}$  is a pivotal quantity, meaning that its distribution does not depend on  $\theta$ . The induced prior measure is classically given by  $\Pi^r(d\theta) \propto \theta^{-1}d\theta$ . Since the multiplication transformation is commutative, the modulus  $\Delta$  is uniformly equal to 1, so that the test statistic that appears in the p-value (corollary 2) is simply  $T(x) = S(x)\theta_0^{-1}f(S(x)\theta_0^{-1})$  where  $\theta_0$  is the value of the parameter under  $H_0$ . For a more general insight into the relationship between Haar invariance and the Fisher pivotal theory, see Eaton and Sudderth (1999).

The theorem 2 assumes that  $\mathcal{G}$ ,  $\mathcal{X}$  and  $\Theta$  are isomorphic. This assumption is relaxed in the corollaries 1 and 2 where the sample  $X$  is replaced by a sufficient statistic  $S(X)$ :  $\mathcal{G}$ ,  $S(\mathcal{X})$  and  $\Theta$  are assumed to be isomorphic. This trick is one of the two classical dimensionality reduction techniques concerning Haar measures applied to statistical problems and somehow restricts the likelihood to belong to the exponential family from Darrois theorem. The second trick consists schematically in replacing  $S(\mathcal{X})$  by the orbit of  $\mathcal{G}$  associated to the observed dataset  $O_{x_0} = \{gx_0 \mid g \in \mathcal{G}\} \subset \mathcal{X}$ . However, the whole set of assumptions that would be involved is more technical, see for example the general assumptions made by Zidek (1969) or Eaton and Sudderth (2002), and not investigated here.

**2.4. Connection to other Bayesian and frequentist reconciliations.** The result, which concerns hypothesis testing, may be related to the different approaches used to reconcile frequentist and Bayesian point estimation somehow and confidence interval especially.

Group invariance applied to invariant inference is the classical frame of such unifications. The Fisherian pivotal theory (Fisher (1956)) is an important contribution mainly to the “frequentist” side and the right Haar measure to the “Bayesian” side. The reconciliation of the two approaches has started with Fraser (1961) and has been deeply studied since then, by Zidek (1969) for example. The most general stage of unification is reached by Eaton and Sudderth (1999). They explicit the central hypothesis of the Fisherian pivotal theory and show under quite standard assumptions in invariance that this hypothesis leads to a procedure which is identical to the Bayesian invariant procedure when using the prior induced by the right Haar measure. Note that they also show (and in a more general manner by Eaton and Sudderth (2002)) that for a Bayesian invariant inference to be admissible (in the sense that there exists no invariant inference whose mean quadratic error is lower for all  $\theta$ ) it has to be obtained from the right Haar prior.

More concretely, the question related to reconciled probability domains is: “Under what assumptions does the following equality hold?”

$$(9) \quad \Pr(\theta \in \mathcal{R}(x)|x) = \Pr(\theta \in \mathcal{R}(x)|\theta)$$

$$\text{i.e.} \quad \int_{\{\theta \in \mathcal{R}(x)\}} d\theta \pi(\theta|x) = \int_{\{x|\theta \in \mathcal{R}(x)\}} dx p(x|\theta)$$



For the equality to hold, each probability needs to be a constant. After Fraser (1961) initial work, Stein (1965) sketched the first conditions of what would be called later Stein's theorem for invariant domains. The part which is common to the different "Stein's theorems" is the following:

*If a domain  $\mathcal{R}(x) \subset \Theta$  satisfies  $\bar{g}\mathcal{R}(x) = \mathcal{R}(g(x))$  with  $\bar{g}\mathcal{R}(x) = \{\bar{g}\theta \mid \theta \in \mathcal{R}(x)\}$ , then under [some invariance assumptions],*

$$\begin{aligned} \Pr(\theta \in \mathcal{R}(x)|x) &= c \quad \forall x \in \mathcal{X} \text{ (Bayesian probability)} \\ \text{and } \Pr(\theta \in \mathcal{R}(x)|\theta) &= c \quad \forall \theta \in \Theta \text{ (frequentist probability)} \end{aligned}$$

One of the simplest set of assumptions found since Stein (1965) is the one of Chang and Villegas (1986). It is relatively close to the one used for our results, presented in the section 2.2.

Our result, mainly holding in the theorem 1, is not a consequence of Stein's theorem because the domain  $\mathcal{R}(x) \subset \Theta$  is not invariant in our case.  $\mathcal{R}(x)$  would be invariant only if  $\theta_0$  was invariant under the transformations group  $\mathcal{G}$ , i.e. if  $\bar{g}\theta_0 = \theta_0$  for all  $\bar{g}$  (this is equivalent to assuming that  $H_0$  is invariant under  $\mathcal{G}$ ). But in the theorem 2, expressed and proved in the appendix 2 and used in the appendix 3 to prove the theorem 1,  $\phi_\theta$  is assumed to be one-to-one for all  $\theta \in \Theta$ , which implies that  $\bar{g}\theta_0 = \theta_0$  is equivalent to  $\bar{g} = e$  (identity function). So the domain  $\mathcal{R}(x) \subset \Theta$  is not invariant in our case and Stein's theorem does not imply the reconciliation result presented in the section 2.2.

The theorem 1 does not answer the previous question, but rather relaxes the form of the domain and accepts a procedure that varies according to the observed dataset  $x_0$  and the value of the parameter  $\theta_0$  under  $H_0$ . It answers to the question: "Under what assumptions and for what domains  $\mathcal{R}$  and  $\mathcal{C}$  does the following equality hold?"

$$(10) \quad \int_{\mathcal{R}(x_0, \theta_0) \subset \Theta} d\theta \pi(\theta|x_0) = \int_{\mathcal{C}(x_0, \theta_0) \subset \mathcal{X}} dx p(x|\theta_0)$$

The domains found take the form

$$\begin{aligned} \mathcal{R}(x_0, \theta_0) &= \{\theta \mid p(x_0|\theta_0) \leq p(x_0|\theta)\} \\ \mathcal{C}(x_0, \theta_0) &= \{x \mid p(x_0|\theta_0)f(x_0) \leq p(x|\theta_0)f(x)\} \end{aligned}$$

where  $f(x)$  is some weighting function, actually given by the inverse of the left prior induced by the underlying group.

### 3. PLR FOR COMPOSITE VS COMPOSITE HYPOTHESES TESTING

Up to this section, the PLR has been only defined in the simple ( $H_0 : \theta = \theta_0$ ) vs composite case, ie according to Dempster (1973)'s first definition.

For the more general hypothesis  $H_0 : \Psi(\theta) = \psi_0$  presented at the end of the section 1.1, Dempster's approach has been generalized by Aitkin (1997), with a modification presented by Aitkin (2010). Namely, Aitkin (2010) proposes to compute  $\Pr(p(x|\theta) < p(x|(\Psi, \Lambda)^{-1}(\psi_0, \Lambda(\theta))) \mid x)$  and details and illustrates some advantages of the method. In the case of  $\Psi(\theta) = \theta$ , it corresponds to Dempster's definition (see page 42 of Aitkin (2010)). The approach of Evans (1997) also carries interesting properties. In particular, a variety of optimality properties for inferences based on relative belief ratios are established in Evans et al. (2006), Evans and Shakhathreh (2008) and Evans and Jang (2011), which include optimal testing properties based on establishing a kind of Bayesian version of the Neyman-Pearson lemma.

However, the hypotheses test case on which they rely is not broad enough for many cases. The purpose of this section is to extend the definition of the PLR to the classical composite vs composite hypotheses test.

Suppose the data models related to the two hypotheses belong to the same parametric family  $\mathcal{P}_\Theta = \{p(\cdot|\theta), \theta \in \Theta\}$ . This assumption can actually be realized for any hypotheses test of parametric models by merging the tested parametric families in a so-called *super-model*. A

composite vs composite hypotheses test consists in choosing among

$$(11) \quad H_0 : \theta \in \Theta_0 \quad H_1 : \theta \in \Theta_1$$

for any domains  $\Theta_0$  and  $\Theta_1$ . We note  $\Pi_0(\cdot)$  and  $\Pi_1(\cdot)$  the prior distributions over  $\Theta_0$  and  $\Theta_1$ .

In this section we propose two extensions of Dempster's approach for this test case. The first extension proposed can be used when the prior under one hypothesis is improper but both posteriors are proper. The second extension, made of two symmetrical probabilities, is the statistics suggested by a new Bayesian-type Neyman-Pearson lemma which also indicates that the LR is a central discrepancy variable.

**3.1. Extensions of the PLR.** In the simple  $\Theta_0 = \{\theta_0\}$  vs composite hypotheses test, the PLR was primarily defined as

$$\text{PLR}(x, \zeta) = \int_{\{\theta_1 | p(x|\theta_1) < \zeta p(x|\theta_0)\}} \Pi_1(d\theta_1|x)$$

In the composite vs composite hypotheses test, a first interesting extension of this concept consists in defining the following statistics:

$$(12) \quad \text{PLR}_{01}(x, \zeta) = \int_{\{(\theta_0, \theta_1) | p(x|\theta_0) < \zeta p(x|\theta_1)\}} \Pi_0(d\theta_1|x) \Pi_1(d\theta_0|x)$$

It is well defined as soon as the posterior distributions are both proper. Since only  $x$  is known, the event  $p(x|\theta_0) < \zeta p(x|\theta_1)$  can be measured only by integrating over all  $\theta_0 \in \Theta_0$  and all  $\theta_1 \in \Theta_1$ . Here we decide to measure it according to the posterior distribution of  $\theta_0$  times the posterior distribution of  $\theta_1$ , which is perfectly allowed.

A second interesting extension of the simple PLR consists in defining the two symmetrical following statistics:

$$(13) \quad \text{PLR}_0(x, \zeta) = \int_{\{(\theta_0, \theta_1) | p(x|\theta_1) < \zeta p(x|\theta_0)\}} \Pi_0(d\theta_0|x) \Pi_1(d\theta_1)$$

$$(14) \quad \text{PLR}_1(x, \zeta) = \int_{\{(\theta_0, \theta_1) | p(x|\theta_0) < \zeta p(x|\theta_1)\}} \Pi_1(d\theta_1|x) \Pi_0(d\theta_0)$$

In the simple vs composite test, note that only  $\text{PLR}_{01}(x, \zeta)$  and  $\text{PLR}_1(x, \zeta)$  are equal to the PLR as defined by Dempster (1973) and can thus be considered as extensions of the PLR. However, given the symmetry of the two hypotheses in a composite vs composite test, the notation  $\text{PLR}_0(x, \zeta)$  will be also necessary in the sequel.

Each quantity has its own definition, interpretation, properties and field of use. We don't investigate interpretation far here, and rather focus on unquestionable properties and results.

$\text{PLR}_{01}(x, \zeta)$  is the only extension of the two which allows for using improper priors. It will be illustrated in the next subsection to test a practical precipitation change, which requires the use of a prior which is too smooth for the other extension to be used.

On the other side, the statistics  $\text{PLR}_1(x, 1)$  is the expectation over the prior under  $H_0$  of the posterior probability under  $H_1$  that the likelihood of  $\theta_0$  is less than the likelihood of  $\theta_1$ , and reciprocally.

$$\text{PLR}_1(x, \zeta) = \mathbb{E}_0[\text{Pr}_1(p(x|\theta_0) < \zeta p(x|\theta_1)|x)]$$

$\text{PLR}_0$  and  $\text{PLR}_1$  will appear as statistics emerging from a more general frame through a Bayesian-type Neyman-Pearson lemma.

Extending the interpretation of the new PLRs in terms of joint probabilities requires the definition of a measure over  $\Theta_0 \times \Theta_1$  given  $x$  and one of the two hypotheses. Such a measure seems to make sense in terms of both mathematics and interpretation but the issue needs to be deepened.

**Remark 1.** *If all subsets defined on the sets  $\Theta_0 \times \mathcal{X}|H_0$  and  $\Theta_1|H_0$  are independent, then the joint measure  $\Pi_{01,0}$  defined over  $\Theta_0 \times \Theta_1 \times \mathcal{X}|H_0$  is equal to:*

$$\Pi_{01,0}(d\theta_0, d\theta_1|x) = \Pi_0(d\theta_0|x) \Pi_1(d\theta_1)$$

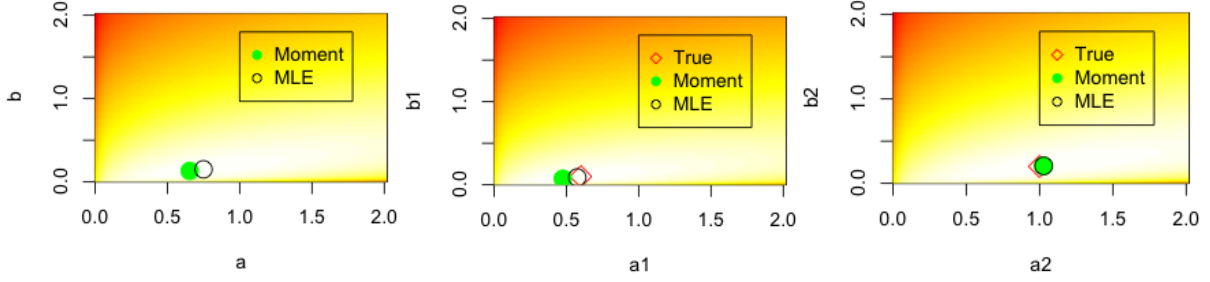


FIGURE 1. Log-likelihood of the dataset under  $H_0$  (left figure) and the dataset under  $H_1$  (center and right figures). Some frequentist estimations of the parameters (circles) are superimposed on the true parameters values (diamond).

for infinitesimal subsets around any  $(\theta_0, \theta_1) \in \Theta_0 \times \Theta_1$ . The same holds when replacing the roles of  $H_0$  and  $H_1$ , and leads to the measure  $\Pi_{01,1}$ :

$$\Pi_{01,1}(d\theta_0, d\theta_1|x) = \Pi_0(d\theta_0)\Pi_1(d\theta_1|x)$$

The proof of the remark stands in the appendix 4. So if we assume that the joint measures exist and that the priors and posteriors are all proper, then the composite PLRs defined in the equations (13) and (14) are probability measures.

**3.2. Example: detection of a change in precipitation in Switzerland.** Let's illustrate  $\text{PLR}_{01}$  defined in the equation (12).

Although the change in temperature in the 20th century is evident at a world scale and in some areas, a potential change in precipitation remains under study. As a simple case, let's consider a single weather station in Switzerland and test whether the statistical properties of the rain frequency have changed.

As recalled for example by Aksoy (2000), daily precipitation amounts are well described by a gamma distribution, characterized by a shape parameter  $a$  and a rate parameter  $b$ . Assume the daily rainfalls  $x_1$  fallen during the five first autumns of the 20th century are i.i.d. with parameters  $a_1$  and  $b_1$ , as well as  $x_2$  during the five last autumns with parameters  $a_2$  and  $b_2$ . The detection of a statistical change consists in testing whether the set of parameters are equal or not:

$$(15) \quad H_0 : (a_1, b_1) = (a_2, b_2) \quad H_1 : (a_1, b_1) \neq (a_2, b_2)$$

Note that the dimension of  $\Theta_0 = \mathbb{R}_{+*}^2$  is less than the dimension of  $\Theta_1 = \mathbb{R}_{+*}^4$ , so that for a regular prior under  $H_1$ ,  $\Pr_1(\theta \in \Theta_0) = 0$ . Borges and Stern (2007) are particularly interested by the behavior of the e-value of the FBST in such cases. Here it simply means that there is one prior  $\pi(a, b)$  under  $H_0$  and the product of two priors  $\pi(a_1, b_1) \times \pi(a_2, b_2)$  under  $H_1$ , to be combined respectively with the likelihood  $p(x_1, x_2|a, b)$  under  $H_0$  and the likelihood  $p(x_1|a_1, b_1) \times p(x_2|a_2, b_2)$  under  $H_1$ .

To enable simple simulations of the posterior distributions under both hypotheses, the conjugate prior (see the compedium by Fink (1997)) of the gamma distribution developed by Miller (1980) is used for  $\pi$ , with hyperparameters that may vary without affecting much the final results. The impact of the prior on the PLR is easy to see from the PLR display as will be explained very shortly. In practice, the prior  $\pi$  is almost improper so that only the  $\text{PLR}_{01}$  defined in equation (12) can be used.

First, simulations roughly corresponding to the observed rainfall are performed. One dataset is simulated under  $H_0$  and another is simulated under some reasonably similar alternative  $H_1$ . The two simulated datasets are characterized by their likelihoods, displayed on the figure 1.

The posterior distribution of each couple  $(a, b)$ ,  $(a_1, b_1)$  and  $(a_2, b_2)$  is separately sampled by a MCMC multivariate slice sampling algorithm (Radford (2003)) implemented in the R package

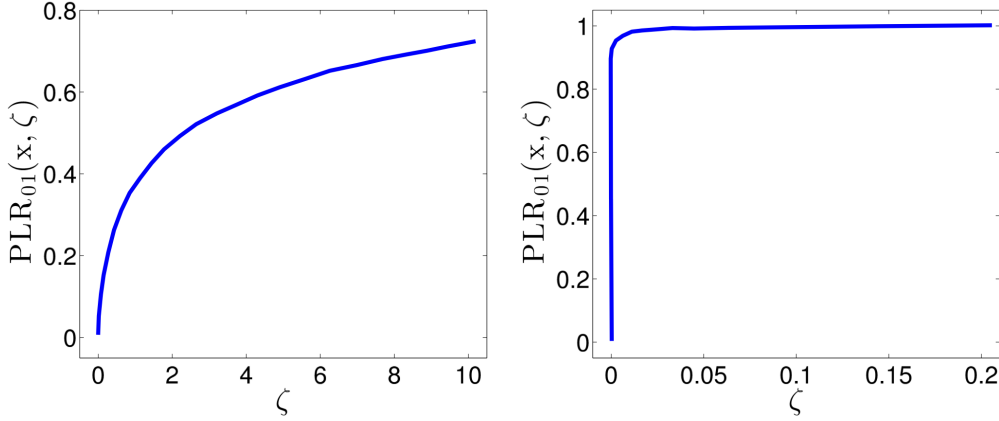


FIGURE 2. PLR obtained from the dataset simulated under  $H_0$  (left) and the dataset simulated under  $H_1$  (right). In practice those are simply the empirical cumulative distributions of the  $LR(x, \theta^{[i]})$  chains. For the  $H_0$  dataset the PLR clearly correctly accepts  $H_0$ , and reciprocally for the  $H_1$  dataset the PLR clearly correctly rejects  $H_0$ : notice the difference of the x axes scales between both simulation cases.

“SamplerCompare” kindly written and provided by Thompson (2012). The PLR is simply computed by ordering the LR obtained for all possible combinations of parameters and counting the fraction which is less than some threshold  $\zeta$  chosen according to the level of evidence wanted in favor of  $H_0$  or  $H_1$ . In practice, the PLR is displayed as a function of  $\zeta$  by simply displaying the empirical cumulative distribution of the LRs. This leads to the figure 2. It can be read for example that for the dataset under  $H_0$ ,  $PLR_{01}(x, 0.1) = 0.08$ , which means that there is an almost null probability that the likelihood under  $H_1$  is more than 10 times greater than the likelihood under  $H_0$ , so that  $H_0$  is (correctly) clearly accepted. Alternatively, for the  $H_1$  dataset,  $PLR_{01}(x, 0.1) = 1.00$ , meaning that there is a probability one that the likelihood under  $H_1$  is more than 10 times greater than the likelihood under  $H_0$ , so that  $H_0$  is (correctly) clearly rejected.

Note that since the GLR indicates the lower bound of the support of the PLR and since the slope of the PLR is infinite there if the likelihood function is smooth enough at its maximum (see section 1.2), the prior exact expression only affects the way  $PLR_{01}(x, \zeta)$  increases as  $\zeta$  departs from  $GLR(x)$ . Here for example the choice of the hyperparameters (among a domain considered as *reasonable*) does not change the conclusion that would be drawn from the PLR displayed on the figure 2.

Switching to the true dataset  $x$ , the PLR is obtained following the same procedure as with the simulated datasets.  $PLR_{01}(x, \zeta)$  is displayed on the figure 3. The graph is –by construction of the simulations– very similar the one obtained for the graph obtained with the data simulated under  $H_0$ . Now,  $PLR_{01}(x, 0.1) = 0.10$  and  $H_0$  can clearly not be rejected, so that no change in the 20th precipitation in Switzerland is detected, which is not surprising to climatologists.

**3.3. Bayesian type Neyman-Pearson lemma.** In the choice of an hypothesis, instead of considering the subset

$$(16) \quad \mathcal{R}^*(x) = \{(\theta_0, \theta_1) \mid p(x|\theta_0) < \zeta p(x|\theta_1)\}$$

one might consider any subset  $\mathcal{R}(x) \subset \Theta_0 \times \Theta_1$ , that may depend on  $x$ . Such a subset could involve a discrepancy variable  $D : \mathcal{X} \times \Theta \mapsto \mathbb{R}$  like in the predictive p-value highlighted by Meng (1994), and take the form “ $D(x, \theta_0) < \zeta D(x, \theta_1)$ ”. The discrepancy variable that appears in the PLR is  $LR(x, \theta)$ .

$\mathcal{R}^*(x)$  defined from the LR test is interesting for hypotheses testing because this set is a somehow classical *hypothesis* rejection set. It is not a fully classical rejection set because it is defined on the parameter space rather than on the observation space, but its characterization

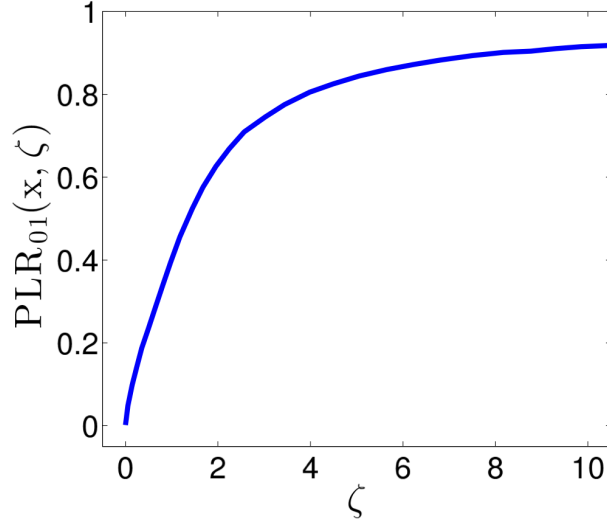


FIGURE 3. PLR obtained from daily autumnal precipitation observed in a weather station in Switzerland from 1900-1905 for the first part of the dataset and 1995-2000 for the second part of the dataset.  $H_0$  cannot be rejected, so that no change in precipitation is detected.

is optimal in the frequentist setting.  $\mathcal{R}^*(x)$  is the set, depending on the dataset  $x$ , of all fixed  $(\theta_0, \theta_1) \in \Theta_0 \times \Theta_1$  such that the likelihood of  $\theta_0$  is less than the likelihood of  $\theta_1$ , which reasonably leads to reject  $H_0$  for this element  $(\theta_0, \theta_1)$ . The same way, one can replace the LR test by any test, ie consider any subset  $\mathcal{R}(x) \subset \Theta_0 \times \Theta_1$  such that for  $(\theta_0, \theta_1) \in \mathcal{R}(x)$ ,  $H_0$  would be decided to be rejected.

With such a phrasing, it may appear natural that the frequentist Neyman-Pearson lemma can be derived in a reciprocal, somehow Bayesian, frame. Note that the Neyman-Pearson lemma can be expressed, as will be the proposition here, symmetrically in the two hypotheses. The symmetry is only broken when adopting the Neyman paradigm which fixes a level for the PFA and deduce the corresponding  $\zeta$  (see section 1.1).

To rederive a Neyman-Pearson lemma one would define the reciprocal notions of “Probability of False Alarm” and “Probability of good Detection”:

$$(17) \quad \text{PFA}_B(\mathcal{R}, x) = \int_{\mathcal{R}(x)} \Pi_0(d\theta_0|x) \Pi_1(d\theta_1)$$

$$(18) \quad \text{PD}_B(\mathcal{R}, x) = \int_{\mathcal{R}(x)} \Pi_1(d\theta_1|x) \Pi_0(d\theta_0)$$

These quantities would also define probability measures if the joint measures exist and if the priors and posteriors are all proper.

Note that these measures can be related to a joint measure with no conditioning over the hypothesis: for any set  $\mathcal{R}(x) \subset \Theta_0 \times \Theta_1$  eventually depending on  $x$ ,

$$\Pr(\mathcal{R}, x) = \Pr(H_0) \Pr(\mathcal{R}|x, H_0) + \Pr(H_1) \Pr(\mathcal{R}|x, H_1)$$

$$\begin{aligned} \text{with } \Pr(\mathcal{R}|x, H_i) &= \int_{\mathcal{R}} \Pi_{01,i}(d\theta_0, d\theta_1|H_i, x) \\ &= \int_{\mathcal{R}} \Pi_i(d\theta_i|x) \Pi_j(d\theta_j) \end{aligned}$$

$$\begin{aligned} \text{so } \Pr(\mathcal{R}|x) &= \Pr(H_0) \int_{\mathcal{R}} \Pi_0(d\theta_0|x) \Pi_1(d\theta_1) + \Pr(H_1) \int_{\mathcal{R}} \Pi_1(d\theta_1|x) \Pi_0(d\theta_0) \\ &= \Pr(H_0) \text{PFA}_0(\mathcal{R}, x) + \Pr(H_1) \text{PD}_1(\mathcal{R}, x) \end{aligned}$$

The Bayesian type probabilities  $\text{PFA}_B$  and  $\text{PD}_B$  add up the same way type I and type II probability errors add up in a frequentist integral. Note also that  $\text{PFA}_B(\bar{\mathcal{R}}, x) = 1 - \text{PFA}_B(\mathcal{R}, x)$  where  $\bar{\mathcal{R}}(x)$  is the set complementary to  $\mathcal{R}(x)$  in  $\Theta_0 \times \Theta_1$ .

Following the underlying idea of the Neyman-Pearson approach, a possibility for choosing  $\mathcal{R}(x)$  consists in maximizing  $\text{PD}_B(\mathcal{R}, x)$  over  $\mathcal{R}(x)$  for a fixed  $\text{PFA}_B(\mathcal{R}, x)$ .

**Proposition 1.** *The subset that maximizes  $\text{PD}_B(\mathcal{R}, x)$  for a fixed value of  $\text{PFA}_B(\mathcal{R}, x)$  is equal to the LR subset  $\mathcal{R}^*(x)$  defined in equation (16). In this case, the “Bayesian PFA and PD” are given by  $\text{PFA}_B(\mathcal{R}^*, x) = 1 - \text{PLR}_0(x, \zeta)$  and  $\text{PD}_B(\mathcal{R}^*, x) = \text{PLR}_1(x, \zeta)$ .*

*Reciprocally, the subset that maximizes  $\text{PFA}_B(\mathcal{R}, x)$  for a fixed value of  $\text{PD}_B(\mathcal{R}, x)$  is equal to  $\bar{\mathcal{R}}^*(x)$ , ie the set which accepts  $H_0$  according to the LR test. In this case,  $\text{PFA}_B(\bar{\mathcal{R}}^*, x) = \text{PLR}_0(x, \zeta)$  and  $\text{PD}_B(\bar{\mathcal{R}}^*, x) = 1 - \text{PLR}_1(x, \zeta)$ .*

As postdata measures (i.e. depending on the observed data), contrary to the predata frequentist PFA and PD, it is therefore informative enough to give  $\text{PLR}_0(x, \zeta)$  and  $\text{PLR}_1(x, \zeta)$  for some value  $\zeta$  of interest. But this is only possible if the priors and posteriors under both hypotheses are proper.

The proof of the proposition follows the proof of the Neyman-Pearson lemma restricted to deterministic tests. It stands in the appendix 5.

#### 4. CONCLUDING GENERAL DISCUSSION ABOUT THE PLR

The PLR introduced by Dempster (1973) in the simple vs composite hypotheses test deserves much attention. It compares the original likelihoods  $p(x|\theta_0)$  and  $p(x|\theta_1)$  by computing the posterior probability that this usual LR test chooses  $H_0$  or  $H_1$ . The PLR is simple, nicely interpretable and coupled with some deep properties. Compared to the classical Bayesian hypotheses tests, first note that unlike the BF, the PLR can be defined even for improper priors, and unlike  $\text{Pr}(H_0|x)$  it does not require the delicate choice of some  $\text{Pr}(H_0)$ . This is crucial in practice as well as in fundamental issues like Lindley’s paradox.

The PLR also turns out to be a very natural alternative to the BF in many aspects. The PLR first compares (the original likelihoods) and then integrates, whereas the BF first integrates and then compares (the marginal likelihoods). In the simple vs composite hypotheses test, considering  $\text{LR}(x, \theta)$  as a random variable for a fixed  $x$ , the PLR is its posterior cumulative distribution (i.e. the probability of a one sided *credible interval*) whereas the BF is its posterior mean *point estimate*. This credible interval vs point estimate duality between the PLR and the BF also translates in decision theory: Hwang et al. (1992) stressed that  $\text{Pr}(H_0|x)$  does not measure evidence, since this is done only through the likelihood, but measures the accuracy of a test by *estimating* the indicator function  $I_{\Theta_0}(\theta)$ . Also note that being the measure of a credible interval, the PLR is also a natural hypotheses test tool which connects postdata (i.e. conditioned upon  $x$ ) hypotheses testing and credible interval inference. This formal equivalence was known to hold for predata inference (a rejection set is equivalent to a confidence interval) and “known” not to hold for postdata inference for usual Bayesian tools (see Lehmann and Romano (2005) and Goutis and Casella (1997)). Tools like the PLR set up this connection.

However, when generalizing the PLR in the section 3.1, most of these dual properties cannot be generalized to the composite vs composite hypotheses test. Instead, a reciprocity between the PLR and the BF exists through a Neyman-Pearson lemma perspective. The second extension of the PLR has been shown in the section 3.3 to be a somehow optimal measure, in that it measures the set that maximizes  $\text{PD}_B$  for a fixed  $\text{PFA}_B$  (Bayesian-type version of the frequentist Neyman-Pearson lemma). Reciprocally, the BF gives a somehow optimal measure, although in the *frequentist* Neyman-Pearson sense, in that it maximizes the average over  $\pi_1$  of  $\text{PD}(\theta_1)$  for a fixed PFA (frequentist classical Neyman-Pearson lemma but for the marginal likelihood and not the original unknown one).

In the simple vs composite hypotheses test, the connection between the PLR (related to credible interval) and the BF (related to point estimate) has been underlined. Another important

connection lies between frequentist and Bayesian type hypotheses tests, namely frequentist p-values and  $\Pr(H_0|x)$  or PLR. This reconciliation quest has been the subject of many debates, including Lindley's paradox in its most simple form (test of the mean of a Gaussian with a uniform prior), which has only been simply reached by the PLR by Dempster (1973). In the section 2.2 we have generalized this reconciliation result to a quite general invariant frame, close to the one used in Stein's theorem, i.e. in a frame under which confidence and credible intervals are equivalent. Note that invariance is also a perspective adopted to develop and evaluate inferences, and in particular to develop new p-values as done recently by Evans and Jang (2010) for example. For the PLR, standard simple invariance properties directly follows from the simple use of the likelihoods.

To conclude on the contribution of this paper, the equivalence between the PLR and a p-value has been proved in a general invariant frame, which nicely connects to the equivalence between confidence and credible domains. This result may contribute to a better understanding of deep and fundamental issues related to both hypotheses testing and parameter estimation, in both frequentist and Bayesian paradigms.

## REFERENCES

- Aitkin, M. (1997). The calibration of p-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood. *Statistics and Computing*, 7:253–261.
- Aitkin, M. (2010). *Statistical inference: an integrated Bayesian / likelihood approach*. Chapman and Hall.
- Aitkin, M., Boys, R. J., and Chadwick, T. (2005). Bayesian point null hypothesis testing via the posterior likelihood ratio. *Statistics and Computing*, 25(3):217–230.
- Aitkin, M., Liu, C. C., and Chadwick, T. (2009). Bayesian model comparison and model averaging for small-area estimation. *Annals of Applied Statistics*, 3(1):199–221.
- Aksoy, H. (2000). Use of gamma distribution in hydrological analysis. *Turk. J. Engin. Environ. Sci.*, 24:419–428.
- Baskurt, Z. and Evans, M. (2013). Hypothesis assessment and inequalities for Bayes factors and relative belief ratios. *Bayesian Analysis*, 8,3:569–590.
- Berger, J. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of P values and evidence (with discussion). *Journal of the American Statistical Association*, 82:112–139.
- Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer-Verlag, 2nd edition.
- Berger, J. O., Brown, L., and Wolpert, R. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *Annals of Statistics*, 22(4):1787–1807.
- Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses (with discussion). *Statistical Science*, 2(3):317–335.
- Bernardo, J. (2011). *Bayesian Statistics 9*, chapter Integrated objective Bayesian estimation and hypothesis testing. Oxford University Press.
- Birnbaum, A. (1962). On the foundation of statistical inference (with discussion). *Journal of the American Statistical Association*, 57(298):269–326.
- Borges, W. and Stern, J. (2007). The rules of logic composition for the Bayesian epistemic e-values. *Logic journal of the IGPL*, 15(5–6):401–420.
- Casella, G. and Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association*, 82(397):106–111.
- Chang, T. and Villegas, C. (1986). On a theorem of Stein relating Bayesian and classical inferences in group models. *The Canadian Journal of Statistics*, 14(4):289–296.
- Darmois, G. (1935). Sur les lois de probabilité à estimation exhaustive. *Compte-Rendu de l'Académie des Sciences de Paris*, 200(1265–1266).
- Dempster, A. P. (1973). The direct use of likelihood for significance testing. In *Proceedings of Conference on Foundational Questions in Statistical Inference*, pages 335–354, Aarhus, Denmark.

- Dempster, A. P. (1997). Commentary on the paper by Murray Aitkin, and on discussion by Mervyn Stone. *Statistics and Computing*, 7(4):265–269.
- Eaton, M. (1989). *Group invariance applications in Statistics*. Regional Conf. Series in Prob. and Stat.
- Eaton, M. (2007). *Multivariate statistics*. Institute of Mathematical Statistics.
- Eaton, M. and Sudderth, W. (1999). Consistency and strong inconsistency of group-invariant predictive inferences. *Bernoulli*, 5(5):833–854.
- Eaton, M. and Sudderth, W. (2002). Group invariant inference and right Haar measure. *Journal of Statistical planning and inference*, 103(1–2):87–99.
- Evans, M. (1997). Bayesian inference procedures derived via the concept of relative surprise. *Communications in Statistics*, 26:1125–1143.
- Evans, M., Guttman, I., and Swartz, T. (2006). Optimality and computations for relative surprise inferences. *Canadian Journal of Statistics*, 34(1):113–129.
- Evans, M. and Jang, G. (2010). Invariant p-values for model checking. *Annals of Statistics*, 38(1):512–525.
- Evans, M. and Jang, G. (2011). Inferences from prior-based loss functions. Technical Report 1104, Dept. of Statistics, U. of Toronto.
- Evans, M. and Shakhatreh, M. (2008). Optimal properties of some Bayesian inferences. *Electronic Journal of Statistics*, 2:1268–1280.
- Fink, D. (1997). A compendium of conjugate priors. Technical report, Montana State University.
- Fisher, R. A. (1973, 1st ed.: 1956). *Statistical methods and scientific inference*. Oliver and Boyd, 3rd edition.
- Fraser, D. A. S. (1961). The fiducial method and invariance. *Biometrika*, 48(3–4):261–280.
- Goutis, C. and Casella, G. (1997). Relationships between post-data accuracy measures. *Annals of the Institute of Statistical Mathematics*, 49(4):711–726.
- Hwang, J., Casella, G., Robert, C., Wells, M., and Farrell, R. (1992). Estimation of accuracy in testing. *Annals of Statistics*, 20(1):490–509.
- Jeffreys, H. (1961). *Theory of probability*. Oxford University Press, 3rd edition.
- Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer, 3rd edition.
- Lindley, D. (1957). A statistical paradox. *Biometrika*, 44(1–2):187–192.
- Meng, X.-L. (1994). Posterior predictive p-values. *Annals of Statistics*, 22(3):1142–1160.
- Miller, R. (1980). Bayesian analysis of the two-parameter Gamma distribution. *Technometrics*, 22(1):65–69.
- Nachbin, L. (1965). *The Haar integral*. Van Nostrand.
- Newton, M. and Raftery, A. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series B*, 56(1):3–48.
- Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, 36:97–131.
- Neyman, J. and Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, 231:289–337.
- Oh, H. and DasGupta, A. (1999). Comparison of the p-value and posterior probability. *Journal of Statistical planning and inference*, 76(1–2):93–107.
- O’Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society*, 57(1):99–138.
- Pereira, C. and Stern, J. (1999). Evidence and credibility: full Bayesian significance test for precise hypotheses. *Entropy*, 1:104–115.
- Radford, N. (2003). Slice sampling. *Annals of Statistics*, 31(3):705–767.
- Robert, C. P. (2007). *The Bayesian choice*. Springer, 2nd edition.
- Robins, J., van der Vaart, A., and Ventura, V. (2000). Asymptotic distribution of p-values in composite null models. *Journal of the American Statistical Association*, 95(452):1143–1156.
- Royall, R. (1997). *Statistical evidence: a likelihood paradigm*. Chapman and Hall / CRC Press.



- Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of p-values for testing precise null hypotheses. *American Statistician*, 55(1):62–71.
- Smith, I. (2010). *Détection d'une source faible : modèles et méthodes statistiques. Application à la détection d'exoplanètes par imagerie directe*. PhD thesis, Université de Nice Sophia-Antipolis.
- Smith, I. and Ferrari, A. (2010). The posterior distribution of the likelihood ratio as a measure of evidence. In *Maxent*.
- Smith, I. and Ferrari, A. (2014). Equivalence between the posterior distribution of the likelihood ratio and a p-value in an invariant frame. *Bayesian Analysis*.
- Stein, C. (1965). Approximation of improper prior measures by prior probability measures. In *Bernoulli, Bayes, Laplace Festschrift*, pages 217–240. Springer-Verlag.
- Thompson, M. (2012). *R package "SamplerCompare"*.
- Tsao, C. A. (2006). A note on Lindley's paradox. *Test*, 15(1):125–139.
- Villegas, C. (1981). Inner statistical inference II. *Annals of Statistics*, 9(4):768–776.
- Zidek, J. (1969). A representation of Bayesian invariant procedures in terms of Haar measure. *Annals of the Institute of Statistical Mathematics*, 21(1):291–308.

## APPENDIX 1: INTRODUCTION TO INVARIANCE IN STATISTICS

For a locally compact Hausdorff group  $\mathcal{G}$ ,  $K(\mathcal{G})$  denotes the class of all continuous real-valued functions on  $\mathcal{G}$  that have compact support. The left invariant Haar measure on  $\mathcal{G}$  is defined as a Radon measure  $H^l$  such that for all  $f \in K(\mathcal{G})$  and all  $g_0 \in \mathcal{G}$ ,

$$\int_{\mathcal{G}} f(g) H^l(dg) = \int_{\mathcal{G}} f(g_0 g) H^l(dg)$$

The right invariant Haar measure  $H^r$  on  $\mathcal{G}$  is defined as  $H^l$  but replacing  $g_0 g$  by  $g g_0$ . For a given group, both Haar measures exist and are unique up to multiplicative constants.

The (right) modulus  $\Delta$  of  $\mathcal{G}$  is the real positive valued function such that if  $H^l$  is a left invariant Haar measure, then for all  $f \in K(\mathcal{G})$  and all  $g_0 \in \mathcal{G}$ ,

$$(19) \quad \int f(g g_0^{-1}) H^l(dg) = \Delta(g_0) \int f(g) H^l(dg)$$

From the unicity of the Haar measure,  $\Delta$  does not depend on the choice of  $H^l$  and is a continuous function such that for all  $g_1, g_2 \in \mathcal{G}$ ,  $\Delta(g_1 g_2) = \Delta(g_1) \Delta(g_2)$ , which implies that  $\Delta(g^{-1}) = \Delta(g)^{-1}$ . Note that for a group  $\mathcal{G}$  the set of all right Haar measures is equal to the set of the left Haar measures if and only if  $\Delta$  is identically equal to 1. This occurs for example when  $\mathcal{G}$  is compact or commutative.

Concerning the Haar measures on the group  $\mathcal{G}$ , the initial definitions and properties imply that if  $H^l$  is a left invariant Haar measure on  $\mathcal{G}$  and  $\Delta$  the modulus of  $\mathcal{G}$  then for all  $f \in K(\mathcal{G})$

$$(20) \quad \int f(g^{-1}) H^l(dg) = \int f(g) \Delta(g)^{-1} H^l(dg)$$

The modulus also enables to relate right and left invariant Haar measures. From the last property, the measure defined by

$$(21) \quad H^r(dg) = \Delta(g)^{-1} H^l(dg)$$

is a right invariant Haar measure on  $\mathcal{G}$ . The same way, if  $H^r$  is a right invariant Haar measure on  $\mathcal{G}$ , then the measure defined by  $H^l(dg) = \Delta(g) H^r(dg)$  is a left invariant Haar measure.

The Haar measure is applied to statistics through the concept of invariance of a data model under a group of transformations. A parametric family  $\mathcal{P}_{\Theta} = \{p(\cdot|\theta), \theta \in \Theta\}$  of densities with respect to any measure  $\mu$  on  $\mathcal{X}$  is said to be invariant under the transformations group  $\mathcal{G}$  if for each  $g \in \mathcal{G}$  there exists a unique  $\theta^* \in \Theta$  such that if the distribution of  $X$  has the density  $p(\cdot|\theta) \in \mathcal{P}_{\Theta}$  then  $Y = gX$  has the density  $p(\cdot|\theta^*) \in \mathcal{P}_{\Theta}$ . This property defines the action of  $\mathcal{G}$  on  $\Theta$ :  $\theta^*$  may simply be denoted  $\theta^* = \bar{g}\theta$  where  $\{\bar{g}, g \in \mathcal{G}\}$  defines a group.

A measure  $\mu$  on  $\mathcal{X}$  is said to be relatively invariant with multiplier  $\chi$  under the group  $\mathcal{G}$  if for all  $f \in K(\mathcal{X})$  and  $g \in \mathcal{G}$

$$(22) \quad \int f(x)\mu(dx) = \chi(g) \int f(gx)\mu(dx)$$

If we assume that both the family of densities and the measure  $\mu$  are respectively invariant and relatively invariant, schematically we get  $p(x|\theta) = \chi(g)p(gx|g\theta)$  for all  $x \in \mathcal{X}, \theta \in \Theta$  and  $g \in \mathcal{G}$ . For more about the connection between such a multiplier and the Jacobian of the transformation that leads to  $gx$  from  $x$ , see for example Berger (1985) or Eaton (2007). Note that the theorem 2 could be formulated differently, by defining the invariance of a probability model, but this phrasing is less common than the invariance of a family of probability densities and this would have entailed a longer presentation.

To shorten the preliminaries and without assuming any knowledge about group theory, we will not refer to group properties like transitivity, orbits... and will concretely simply assume that  $\Theta$  and  $\mathcal{G}$  are isomorphic. More precisely, we will assume that the transformation  $\phi_\theta : \mathcal{G} \mapsto \Theta$  with  $\phi_\theta(g) = g\theta$  is one-to-one whatever  $\theta \in \Theta$ . The right Haar prior on  $\Theta$  is to be induced from the right Haar measure  $H^r$  on  $\mathcal{G}$  and the action of  $\mathcal{G}$  on  $\Theta$ . From the frame chosen, the right Haar prior  $\Pi_a^r$  is simply defined by  $\Pi_a^r = H^r(\phi_a^{-1})$ , with  $a \in \Theta$ . As shown in Villegas (1981), it turns out that the measure  $\Pi_a^r$  actually does not depend on  $a$ . The induced prior is therefore unique for a fixed  $H^r$  and noted  $\Pi^r$ .  $\Pi^r = H^r(\phi_a^{-1})$  means that for any measurable subset  $A \subset \Theta$ ,  $\Pi^r(A) = H^r(\phi_a^{-1}A)$  with  $\phi_a^{-1}A = \{\phi_a^{-1}\theta | \theta \in A\}$ . Note that a subset  $A = d\theta$  denotes an infinitesimal subset centered around  $\theta$ , where  $\theta$  is implicit.  $\Pi^r$  can be normalized into a probability measure if and only if the group  $\mathcal{G}$  is compact, and in this case we can go back to the usual notation  $\Pi^r(A) = \Pr(\theta \in A)$  where the measure  $\Pi^r$  is implicit in  $\Pr(\cdot)$ .

Finally, from the data model density  $p(\cdot|\theta)$  and the prior  $\Pi^r$ , the posterior measure  $\Pi_x^r$  on  $\Theta$  is classically defined by

$$(23) \quad \Pi_x^r(B) = \frac{\int_B p(x|\theta)\Pi^r(d\theta)}{m(x)} \quad \text{for all } B \subset \Theta$$

$$\text{with } m(x) = \int p(x|\theta)\Pi^r(d\theta)$$

where the marginal  $m(x)$  density of  $x$  is always assumed to be finite, so that  $\Pi_x^r$  defines a probability measure even if  $\Pi^r$  does not. Then the posterior probability of an event is denoted by  $\Pr(\cdot|x)$ , meaning  $\Pr(\theta \in B|x) = \Pi_x^r(B)$ .

## APPENDIX 2: GENERAL THEOREM AND ITS PROOF

**Theorem 2.** Call  $\mathcal{P}_\Theta = \{p(\cdot|\theta), \theta \in \Theta\}$  a family of probability densities with respect to a measure  $\mu^r$  on  $\mathcal{X}$ , specified later, and call  $\mathcal{G}$  a group acting on  $\mathcal{X}$ . Assume that  $\mathcal{P}_\Theta$  is invariant under the action of the group  $\mathcal{G}$  on  $\mathcal{X}$  and note  $\bar{g}\theta$  the induced action of the element  $g \in \mathcal{G}$  on the element  $\theta \in \Theta$ . Call  $H^r$  any right Haar measure of  $\mathcal{G}$  and define the transformations  $\phi_\theta$  (for  $\theta \in \Theta$ ) and  $\phi_x$  (for  $x \in \mathcal{X}$ ) by

$$(24) \quad \begin{array}{ll} \phi_\theta : \bar{\mathcal{G}} & \mapsto \Theta \\ \bar{g} & \mapsto \bar{g}\theta \end{array} \quad \begin{array}{ll} \phi_x : \mathcal{G} & \mapsto \mathcal{X} \\ g & \mapsto gx \end{array}$$

Assume that

- (1)  $\phi_\theta$  is one-to-one for all  $\theta \in \Theta$  and  $\phi_x$  is one-to-one for all  $x \in \mathcal{X}$ .
- (2) The prior measure  $\Pi^r$  on  $\Theta$  is the measure induced by  $H^r$  via  $\phi_\theta$  and the measure  $\mu^r$  on  $\mathcal{X}$  is the measure induced by  $H^r$  via  $\phi_x$ :  $\Pi^r = H^r(\phi_\theta^{-1})$  and  $\mu^r = H^r(\phi_x^{-1})$ .
- (3) The marginal density of  $x$  is finite, so that the posterior measure  $\Pi_x^r$  on  $\Theta$ , classically defined by the equation (23), defines the posterior probability  $\Pr(\cdot|x_0)$ .

Then, the PLR defined by the equation (4) can be reexpressed, for any  $\zeta > 0$  and any  $c \in \mathcal{X}$ , as the frequentist integral:

$$(25) \quad \text{PLR}(x_0, \zeta) = \Pr\left( p(x_0|\theta_0)\Delta(\phi_{x_0}^{-1}c) \leq \zeta p(x|\theta_0)\Delta(\phi_x^{-1}c) \mid \theta_0 \right)$$

where  $\Delta$  is the modulus of the group  $\mathcal{G}$ , as defined in the equation (19), and in practice  $x_0 \in \mathcal{X}$  is the observed data and  $\theta_0 \in \Theta$  the parameter value under the null hypothesis.

Note that as seen in the previous appendix, the measures  $\mu$  and  $\Pi^r$  defined in the theorem 2 do not depend on the choice of  $\theta \in \Theta$  and  $x \in \mathcal{X}$  in the functions  $\phi_\theta$  and  $\phi_x$ . In order to clarify the proof, we note  $a$  instead of  $x$  and  $b$  instead of  $\theta$  in the following. We shall make use of the following lemma:

**Lemma 1.** *The measures  $\mu$  on  $\mathcal{X}$  and  $\Pi^r$  on  $\Theta$  induced above by the right Haar measure  $H^r$  on  $\mathcal{G}$  are relatively invariant with modulus  $\Delta^{-1}$ .*

*Proof.*

$$\begin{aligned} \int f(g_0x)\mu(dx) &= \int f(g_0x)H^r\phi_a^{-1}(dx) \text{ (Def. of } \mu \text{ in the Cond. of Th. 2)} \\ &= \int f(g_0\phi_ag)H^r(dg) \text{ (transformation } g = \phi_a^{-1}x) \\ &= \int f(g_0ga)\Delta(g)^{-1}H^l(dg) \text{ (Def. of } \phi_a \text{ and prop. eq. (21))} \\ &= \Delta(g_0) \int f(g_0ga)\Delta(g_0g)^{-1}H^l(dg) \text{ (Multiplicity prop. of } \Delta) \\ &= \Delta(g_0) \int f(ga)\Delta(g)^{-1}H^l(dg) \text{ (} H^l \text{ left invariant)} \\ &= \Delta(g_0) \int f(x)\mu(dx) \text{ (previous computation made in reverse order)} \end{aligned}$$

This also implies that a Haar prior induced as in the theorem 2, i.e. from a right invariant Haar measure on  $\mathcal{G}$ , is relatively invariant.

$$\begin{aligned} \text{PLR}(x_0, \zeta) &= \Pr\left( p(x_0|\theta_0) \leq \zeta p(x_0|\theta) \mid x_0 \right) \\ (26) \quad &= \frac{1}{m(x_0)} \int_{\{\theta|p(x_0|\theta_0) \leq \zeta p(x_0|\theta)\}} p(x_0|\theta)\Pi^r(d\theta) \\ &= \frac{1}{m(x_0)} \int_{\{\theta|p(x_0|\theta_0) \leq \zeta p(x_0|\theta)\}} p(x_0|\theta)H^r(\phi_b^{-1}(d\theta)) \text{ (Def. } \Pi^r \text{ in the Cond. of Th. 2)} \\ &= \frac{1}{m(x_0)} \int_{\{g|p(x_0|\theta_0) \leq \zeta p(x_0|\phi_bg)\}} p(x_0|\phi_bg)H^r(dg) \text{ (} g = \phi_b^{-1}\theta) \\ &= \frac{1}{m(x_0)} \int_{\{g|p(x_0|\theta_0) \leq \zeta p(x_0|gb)\}} p(x_0|gb)H^r(dg) \text{ (Def. } \phi_\theta \text{ eq. (24))} \\ &= \frac{1}{m(x_0)} \int_{\{g|p(x_0|\theta_0) \leq \zeta p(x_0|gb)\}} p(x_0|gb)\Delta(g)^{-1}H^l(dg) \text{ (Prop. eq. (21))} \\ &= \frac{1}{m(x_0)} \int_{\{g|p(x_0|\theta_0) \leq \zeta p(x_0|g^{-1}b)\}} p(x_0|g^{-1}b)H^l(dg) \text{ (Prop. eq. (20))} \end{aligned}$$

But according to lemma 1,  $\mu$  is relatively invariant with modulus  $\Delta^{-1}$ . Since the density family is invariant,

$$p(x|\theta) = \Delta(g)^{-1}p(gx|g\theta) \text{ for all } x \in \mathcal{X}, \theta \in \Theta, g \in \mathcal{G}$$

i.e.

$$p(x|g^{-1}\theta') = \Delta(g)^{-1}p(gx|\theta') \text{ for all } x \in \mathcal{X}, \theta' \in \Theta, g \in \mathcal{G}$$

Then,

$$\begin{aligned} \text{PLR}(x_0, \zeta) &= \frac{1}{m(x_0)} \int_{\{g: p(x_0|\theta_0) \leq \zeta p(gx_0|b)\Delta(g)^{-1}\}} \Delta(g)^{-1} p(gx_0|b) H^l(dg) \\ &= \frac{1}{m(x_0)} \int_{\{g: p(x_0|\theta_0) \leq \zeta p(gx_0|b)\Delta(g)^{-1}\}} p(gx_0|b) H^r(dg) \text{ (Prop. eq. (21))} \\ &= \frac{1}{m(x_0)} \int_{\{g: p(x_0|\theta_0) \leq \zeta p(gx_0|b)\Delta(g)^{-1}\}} p(gx_0|b) \mu(\phi_a(dg)) \text{ (Def. } \mu) \end{aligned}$$

It can be noticed that the equation (26) depends neither on  $a \in \mathcal{X}$  nor on  $b \in \Theta$ . Choose now for simplicity  $a = x_0$ . Then, making the transformation  $x = \phi_{x_0}g = gx_0$ ,

$$\text{PLR}(x_0, \zeta) = \frac{1}{m(x_0)} \int_{\{x: p(x_0|\theta_0) \leq \zeta p(x|b)\Delta(\phi_{x_0}^{-1}x)^{-1}\}} p(x|b) \mu(dx)$$

By a similar computation we get the expression of the marginal density of  $X$  evaluated at  $x_0$ :

$$m(x_0) = \int p(x_0|\theta) \Pi^r(d\theta) = \int p(x|b) \mu(dx) = 1$$

The marginal density of  $X$  is constant, the same way the frequentist risk of an invariant estimator does not depend on  $\theta$ . So

$$\text{PLR}(x_0, \zeta) = \int_{\{x: p(x_0|\theta_0) \leq \zeta p(x|b)\Delta(\phi_{x_0}^{-1}x)^{-1}\}} p(x|b) \mu(dx)$$

In order to get a form closer to a p-value, we choose from now  $b = \theta_0$  and note that for *any*  $c \in \mathcal{X}$ ,

$$(27) \quad \Delta(\phi_{x_0}^{-1}x) = \frac{\Delta(\phi_c^{-1}x)}{\Delta(\phi_c^{-1}x_0)}$$

because if we note

$$\begin{aligned} g &= \phi_{x_0}^{-1}x \\ g_1 &= \phi_c^{-1}x \\ g_2 &= \phi_c^{-1}x_0 \end{aligned}$$

then on one side  $gx_0 = x$  and on the other  $g_1(g_2^{-1}x_0) = g_1c = x$  so that

$$\begin{aligned} gx_0 &= (g_1g_2^{-1})x_0 \\ \text{so } \phi_{x_0}g &= \phi_{x_0}(g_1g_2^{-1}) \\ \text{so } g &= g_1g_2^{-1} \text{ (}\phi_a \text{ is one-to-one)} \\ \text{so } \Delta(g) &= \frac{\Delta(g_1)}{\Delta(g_2)} \text{ (Prop. of } \Delta) \end{aligned}$$

Finally, for any  $c \in \mathcal{X}$

$$(28) \quad \text{PLR}(x_0, \zeta) = \int_{\left\{x \mid \frac{p(x_0|\theta_0)}{\Delta(\phi_c^{-1}x_0)} \leq \zeta \frac{p(x|\theta_0)}{\Delta(\phi_c^{-1}x)}\right\}} p(x|\theta_0) \mu(dx)$$

It is also interesting to note that

$$(29) \quad \begin{aligned} \phi_a^{-1}b &= (\phi_b^{-1}a)^{-1} \\ \text{since } g &= \phi_a^{-1}b \Rightarrow ga = b \Rightarrow a = g^{-1}b \Rightarrow g^{-1} = \phi_b^{-1}a \end{aligned}$$

so that the same way we have

$$\begin{aligned} \text{PLR}(x_0, \zeta) &= \int_{\{x|p(x_0|\theta_0)\Delta(\phi_{x_0}^{-1}c) \leq \zeta p(x|\theta_0)\Delta(\phi_x^{-1}c)\}} p(x|\theta_0)\mu(dx) \\ &= \Pr\left(p(x_0|\theta_0)\Delta(\phi_{x_0}^{-1}c) \leq \zeta p(x|\theta_0)\Delta(\phi_x^{-1}c) \middle| \theta_0\right) \end{aligned}$$

□

### APPENDIX 3: PROOF OF THE THEOREM 1 AND COROLLARIES 1, 2

The theorem 1 is a corollary of the theorem 2 presented and proved in the previous appendix: the theorem 2 can be reexpressed more simply by assuming that the likelihood family and the induced Haar measures are absolutely continuous with respect to the Lebesgue measure.

*Proof of the theorem 1.* The proof only consists of reexpressing the domains of integration because the integrands expression are not functions of the use of the decomposition of the measures over some other measures ( $\mu$  or Lesbesgue). The proof even actually only consists of reexpressing the domain of integration of the p-value because the domain of integration of the PLR does not depend on the density over  $\mathcal{X}$  used since the domain of integration is a subset of  $\Theta$ , not  $\mathcal{X}$ .

If we note  $p^\mu(\cdot|\theta)$  the density with respect to the induced Haar measure  $\mu^r$  and  $p(\cdot|\theta)$  the density with respect to the Lebesgue measure, we have by definition

$$\begin{aligned} P(dx|\theta) &= p^\mu(x|\theta)\mu^r(dx) = p^\mu(x|\theta)\pi^r(x)dx \quad \text{and} \quad P(dx|\theta) = p(x|\theta)dx \\ \text{and so} \quad p^\mu(x|\theta) &= \frac{p(x|\theta)}{\pi^r(x)} \end{aligned}$$

On the other side the modulus  $\Delta$  can also be reexpressed as a function of the induced prior densities  $\pi^l(x)$  and  $\pi^r(x)$ . From the equations (29) and (21),

$$\Delta(\phi_x^{-1}c) = \Delta(\phi_c^{-1}x)^{-1} = \frac{H^l(d\phi_c^{-1}x)}{H^r(d\phi_c^{-1}x)} = \frac{\mu^r(dx)}{\mu^l(dx)} = \frac{\pi^r(x)}{\pi^l(x)}$$

Combining these two results we get

$$p^\mu(x|\theta)\Delta(\phi_x^{-1}c) = \frac{p(x|\theta)}{\pi^l(x)}$$

□

*Proof of the corollary 1.*

$$\begin{aligned} \text{PLR}(x_0, \zeta) &= \Pr\left((p_{X|\theta_0}(x_0) \leq \zeta p_{X|\theta}(x_0)) \middle| x_0\right) \\ &= \Pr\left(p_{X|S(X)}(x_0|S(x_0)) p_{S(X)|\theta_0}(S(x_0)) \leq \zeta p_{X|S(X)}(x_0|S(x_0)) p_{S(X)|\theta}(S(x_0)) \middle| x_0\right) \end{aligned}$$

because since  $S(x)$  is a function of  $x$ ,  $p_{X|\theta}(x) = p_{X,S(X)|\theta}(x, S(x))$  and since in addition  $S(X)$  is a sufficient statistic of  $X$ ,

$$\begin{aligned} p_{X|\theta}(x) &= p_{X|S(X),\theta}(x|S(x)) p_{S(X)|\theta}(S(x)) \\ (30) \quad &= p_{X|S(X)}(x|S(x)) p_{S(X)|\theta}(S(x)) \end{aligned}$$

Simplifying the densities which do not depend on  $\theta$ ,

$$\begin{aligned} \text{PLR}(x_0, \zeta) &= \Pr\left(p_{S(X)|\theta_0}(S(x_0)) \leq \zeta p_{S(X)|\theta_0}(S(x_0)) \middle| S(x_0)\right) \text{ if } p_{X|S(X)}(x_0|S(x_0)) > 0 \\ &= \Pr\left(p_{S(X)|\theta_0}(S(x_0))(\pi^l(S(x_0)))^{-1} \leq \zeta p_{S(X)|\theta_0}(S(x_0))(\pi^l(S(x_0)))^{-1} \middle| \theta_0\right) \quad (\text{Th. 1}) \end{aligned}$$

□

*Proof of the corollary 2.* First reexpress the PLR under the conditions of the theorem 1 by using a cumulative distribution. Note  $T(x)$  the statistic:

$$T(x) = p_{S(X)|\theta_0}(S(x))(\pi^l(S(x)))^{-1}$$

Seen as a random variable, the dataset  $x$  induces the random variable  $T(X)$  the same way the statistic  $S(x)$  induced  $S(X)$ . Note  $F_{T(X)|\theta_0}$  the cumulative distribution of  $T(X)$  under the null hypothesis:

$$F_{T(X)|\theta_0}(\zeta) = \Pr(T(x) \leq \zeta | \theta_0)$$

Starting from the theorem 1, the PLR can be reexpressed as

$$\text{PLR}(x_0, \zeta) = 1 - F_{T(X)|\theta_0}(\zeta^{-1}T(x_0))$$

In particular, for a threshold  $\zeta = 1$ , one can directly notice that the PLR is equal to the  $p$ -value defined for the GLR by the equation (2), but now instead associated to the test statistic  $T(x)$ .

Also note that the frequentist test corresponding to the PLR is then given, for any threshold  $\lambda > 0$ , by

$$\text{Reject } H_0 \text{ if } p_S(S(x)|\theta_0) (\pi^l(S(x)))^{-1} \leq \lambda$$

□

#### APPENDIX 4: PROOF OF THE REMARK 1

If there exists a joint measure over  $\Theta_0 \times \Theta_1 \times \mathcal{X} | H_0$  and if the events defined on the sets  $\Theta_0 \times \mathcal{X} | H_0$  and  $\Theta_1 | H_0$  are independent, then

$$P(d\theta_0, d\theta_1, x | H_0) = P(d\theta_0, x | H_0)P(d\theta_1 | H_0)$$

This can be reformulated using the standard previous notations:

$$P(d\theta_0, d\theta_1, x | H_0) = \Pi_0(d\theta_0 | x) m_0(x) \Pi_1(d\theta_1)$$

Then, noting  $\Pi_{01,0}(\cdot | x) = P(\cdot | H_0, x)$ ,

$$\begin{aligned} \Pi_{01,0}(d\theta_0, d\theta_1 | x) &= \frac{P(d\theta_0, d\theta_1, x | H_0)}{\int P(d\theta_0, d\theta_1, x | H_0)} \\ &= \frac{m_0(x) \Pi_0(d\theta_0 | x) \Pi_1(d\theta_1)}{m_0(x)} \\ &= \Pi_0(d\theta_0 | x) \Pi_1(d\theta_1) \end{aligned}$$

#### APPENDIX 5: PROOF OF THE PROPOSITION 1

Recall that the set  $\mathcal{R}^*(x)$  defined in equation (16) is the LR set that rejects  $H_0$ , and upon which is defined  $\text{PLR}_1$  in equation (14). Call  $\text{PFA}_B(\mathcal{R}^*, x)$  and  $\text{PD}_B(\mathcal{R}^*, x)$  the associated integrals defined in equations (17) and (18). Call  $\mathcal{R}(x) \subset \Theta_0 \times \Theta_1$  any other set and  $\text{PFA}_B(\mathcal{R}, x)$  and  $\text{PD}_B(\mathcal{R}, x)$  its associated integrals.

The goal is to show that  $\text{PFA}_B(\mathcal{R}, x) \leq \text{PFA}_B(\mathcal{R}^*, x)$  implies that  $\text{PD}_B(\mathcal{R}, x) \leq \text{PD}_B(\mathcal{R}^*, x)$  for any test set  $\mathcal{R}$ . The fact that  $\text{PD}_B(\mathcal{R}, x) \leq \text{PD}_B(\mathcal{R}^*, x)$  implies that  $\text{PFA}_B(\mathcal{R}, x) \leq \text{PFA}_B(\mathcal{R}^*, x)$  is shown in a reciprocal way.

One can check that the following inequality holds for all  $x \in \mathcal{X}$ , all  $\theta_0 \in \Theta_0$  and all  $\theta_1 \in \Theta_1$ :

$$(I_{\mathcal{R}^*(x)}(\theta_0, \theta_1) - I_{\mathcal{R}(x)}(\theta_0, \theta_1)) (p(x|\theta_0) - \zeta p(x|\theta_1)) \leq 0$$

Since the inequality is true for all  $x, \theta_0$  and  $\theta_1$ , we can multiply the left hand side by any positive term and integrate over  $\Theta_0 \times \Theta_1$ . This implies in particular:

$$\int_{\Theta_0} \int_{\Theta_1} \frac{\Pi_0(d\theta_0)}{m_0(x)} \frac{\Pi_1(d\theta_1)}{m_1(x)} (I_{\mathcal{R}^*(x)}(\theta_0, \theta_1) - I_{\mathcal{R}(x)}(\theta_0, \theta_1)) (p(x|\theta_0) - \zeta p(x|\theta_1)) \leq 0$$

But since  $\Pi_i(d\theta_i|x) = \Pi_i(d\theta_i)p(x|\theta_i)m_i(x)^{-1}$  for  $i = 0, 1$ , this implies

$$\begin{aligned} & \frac{1}{m_1(x)} \int_{\Theta_0} \int_{\Theta_1} \Pi_0(d\theta_0|x) \Pi_1(d\theta_1) (I_{\mathcal{R}^*(x)}(\theta_0, \theta_1) - I_{\mathcal{R}(x)}(\theta_0, \theta_1)) \\ & - \frac{\zeta}{m_0(x)} \int_{\Theta_0} \int_{\Theta_1} \Pi_1(d\theta_1|x) \Pi_0(d\theta_0) (I_{\mathcal{R}^*(x)}(\theta_0, \theta_1) - I_{\mathcal{R}(x)}(\theta_0, \theta_1)) \leq 0 \end{aligned}$$

where we recognize  $\text{PFA}_B$  and  $\text{PD}_B$  as defined in equations (17) and (18):

$$\frac{\text{PFA}_B(\mathcal{R}^*(x)) - \text{PFA}_B(\mathcal{R}(x))}{m_1(x)} - \zeta \frac{\text{PD}_B(\mathcal{R}^*(x)) - \text{PD}_B(\mathcal{R}(x))}{m_0(x)} \leq 0$$

Therefore we finally have

$$\zeta \frac{\text{PD}_B(\mathcal{R}(x)) - \text{PD}_B(\mathcal{R}^*(x))}{m_0(x)} \leq \frac{\text{PFA}_B(\mathcal{R}(x)) - \text{PFA}_B(\mathcal{R}^*(x))}{m_1(x)}$$

from which we conclude the final implication

$$\text{PFA}_B(\mathcal{R}(x)) \leq \text{PFA}_B(\mathcal{R}^*(x)) \Rightarrow \text{PD}_B(\mathcal{R}(x)) \leq \text{PD}_B(\mathcal{R}^*(x))$$

The fact that  $\text{PD}_B(\mathcal{R}, x) \leq \text{PD}_B(\bar{\mathcal{R}}^*, x)$  implies that  $\text{PFA}_B(\mathcal{R}, x) \leq \text{PFA}_B(\bar{\mathcal{R}}^*, x)$  is shown in a reciprocal way.